Baoflowin502 at MAHED Shared Task: Text-based Hate and Hope Speech Classification

Nguyen Minh Bao
University of Information Technology
23520123@gm.uit.edu.vn

Dang Van Thin
University of Information Technology

Abstract

This paper presents Arabic hate and hope speech classification using pre-trained language models and advanced data augmentation techniques. We evaluate multiple Arabic BERT variants on 6,889 Arabic text samples labeled as hate speech, hope speech, or not-applicable. Data augmentation through back-translation and LLM-based data generation using few-shot prompting significantly improves performance across all models. We establish strong baselines using ensemble bagging XGBoost alongside traditional machine learning approaches. CAMeLBERT with data augmentation achieves the best Macro-F1 of 0.6868, demonstrating the effectiveness of Arabic-specific models combined with modern augmentation strategies for hate speech detection speech detection.

1 Introduction

The proliferation of social media has transformed communication while significantly accelerating the spread of hate speech and harmful content. In this competition (Bao, 2025) (Zaghouani et al., 2025) (Zaghouani et al., 2024), we tackle the Arabic hate–hope–neutral speech classification problem — a challenging NLP task due to Arabic's morphological richness, diverse dialects, and right-to-left script. These linguistic complexities, along with subtle cultural and contextual cues, make model development more difficult than for high-resource languages like English (Wahdan et al., 2024; Elnagar et al., 2020).

We conduct a systematic evaluation of three leading Arabic-specific BERT variants — AraBERTv2 (Antoun et al.), CAMeLBERT (Inoue et al., 2021), and MARBERT (Abdul-Mageed et al., 2021) — chosen for their complementary strengths in handling formal, morphologically complex, and multi-dialectal Arabic text. To mitigate data scarcity, we adopt a dual augmentation strategy: multi-hop back-translation to generate natural paraphrases and LLM-based few-shot prompting (Kim et al., 2024) to produce contextually coherent synthetic examples. For comparison, we also implement strong traditional baselines using ensemble bagging XGBoost on contextual embeddings.

Our contributions include: a systematic benchmark of high-performing Arabic BERT variants for hate–hope speech classification, a tailored augmentation pipeline for Arabic text that combines cross-lingual and generative approaches, and the development of a hard voting ensemble method that leverages the complementary abilities of multiple models, achieving consistent performance improvements and advancing Arabic NLP for content moderation applications.

2 Related Work

2.1 Text Preprocessing for Arabic Social Media

Effective preprocessing involves converting emojis to textual representations using comprehensive emoji-to-text dictionaries to preserve emotional context essential for hate and hope sentiment analysis. Text normalization through tokenization ensures consistent input representation, handling variations in spelling, punctuation, and formatting commonly found in social media posts. Additional preprocessing includes URL removal, mention cleaning, and Arabic text standardization to optimize model performance while preserving linguistically relevant information for classification tasks.

2.2 Data Augmentation Strategies

Data augmentation addresses low-resource scenarios through two sophisticated techniques. Back-translation leverages machine translation systems to generate paraphrases by translating text through intermediate languages and back to the

Proceedings of The Third Arabic Natural Language Processing Conference, pages 665–669

source language, creating diverse training examples while preserving semantic meaning. Large language model-based data generation using few-shot prompting provides contextually appropriate training examples by leveraging in-context learning capabilities with representative sample prompts. The combination of back-translation and LLM-based few-shot generation provides complementary benefits, addressing different aspects of data scarcity while ensuring high-quality augmented datasets.

2.3 Arabic Language Model Selection

Through comprehensive literature survey, we selected three prominent Arabic-specific BERT variants demonstrating superior performance in Arabic NLP tasks: AraBERTv2 (comprehensive Arabic BERT pre-trained on large-scale Arabic corpora), CAMeLBERT (advanced model with optimized architecture for Arabic morphological features), and MARBERT (multi-dialectal Arabic specialist for diverse text processing). This focused selection ensures robust evaluation of the most established Arabic language models for hate and hope speech classification tasks.

3 Methodology

3.1 Dataset Description

Our experimental dataset comprises 6,889 Arabic text samples systematically collected from diverse social media platforms including Twitter, Facebook, and regional Arabic forums. The dataset encompasses three distinct classification categories: **Hate Speech** samples (2,296 instances, 33.3%) containing explicit or implicit expressions of hatred and discrimination targeting individuals or groups; **Hope Speech** samples (2,301 instances, 33.4%) promoting positive values, social inclusion, and constructive dialogue; and **Not-applicable** samples (2,292 instances, 33.3%) representing neutral content that does not clearly fall into either category. The balanced class distribution provides a solid methodological foundation for robust model training, while geographic diversity across different Arab-speaking regions ensures linguistic representativeness.

3.2 Data Preprocessing Pipeline

We implement a comprehensive preprocessing pipeline specifically tailored for Arabic social media text. The process includes emoji replacement using extensive multilingual dictionaries containing over 3,000 mappings to preserve emotional context crucial for sentiment classification; URL detection and removal via robust regular expressions while preserving adjacent contextual information; Arabic text normalization addressing script-specific challenges including variant letter forms and punctuation standardization; tokenization using NLTK's Arabic-specific algorithms enhanced with custom rules for morphological patterns; and systematic mention removal to reduce person-specific bias while maintaining relevant surrounding context.

3.3 Data Augmentation Strategies

We employ three complementary augmentation strategies with a 50% augmentation ratio to balance dataset expansion with computational efficiency:

3.3.1 Multi-hop Back-Translation

Multi-step translation process using Google Translate API following Arabic $\rightarrow$ English $\rightarrow$ French $\rightarrow$ Arabic sequence. This approach introduces natural linguistic variations through different language typologies while preserving semantic content. Quality control includes automatic filtering of translation artifacts and semantic similarity verification using multilingual embeddings.

3.3.2 LLM-based Few-Shot Data Generation

Leveraging GPT-4 with carefully designed prompting strategies providing 3-5 representative examples per class. The methodology includes explicit instructions for maintaining dialectal authenticity, appropriate emotional intensity, and realistic social media communication patterns. Quality assurance involves automated toxicity filtering and semantic coherence verification.

3.3.3 Controlled Lexical Substitution

Systematic replacement using Arabic WordNet and curated synonym dictionaries, selectively targeting non-key terms to introduce lexical diversity without altering core semantic meaning. The process incorporates POS tagging and NER to preserve proper nouns and category-specific terminology essential for classification accuracy.

4 Model Architecture and Experimental Setup

4.1 Pre-trained Language Models

We evaluate three prominent Arabic-specific BERT variants based on comprehensive literature survey:

AraBERTv2 (comprehensive Arabic BERT pretrained on large-scale Arabic corpora), CAMeLBERT (advanced model with optimized architecture for Arabic morphological features), and MARBERT (multi-dialectal Arabic specialist for diverse text processing). This focused selection ensures robust evaluation of the most established Arabic language models for hate and hope speech classification tasks.

4.2 Hard Voting Ensemble Method

To leverage the complementary strengths of individual Arabic BERT models, we implement a hard voting ensemble combining predictions from AraBERTv2, CAMeLBERT, and MARBERT. In this ensemble approach, each model independently processes the input text and generates predictions for the three classes (Hate, Hope, Not-applicable). The final prediction is determined through majority voting, where the class receiving the most votes across the three models is selected as the ensemble output. In cases of tie situations, we implement a confidence-based tie-breaking mechanism using the model with the highest prediction probability. This hard voting strategy capitalizes on the diverse strengths of each Arabic BERT variant: AraBERTv2's comprehensive Arabic coverage, CAMeLBERT's morphological optimization, and MARBERT's dialectal expertise, potentially improving overall classification robustness and accuracy.

4.3 Traditional Machine Learning Baselines

For a comprehensive performance comparison, we establish strong traditional machine learning baselines. Specifically, we implement an ensemble bagging XGBoost classifier that operates on vector embeddings extracted from the AraBERTv2 model. By combining the representational power of contextualized AraBERTv2 embeddings with XGBoost's gradient boosting capabilities, this setup effectively captures both semantic and lexical patterns present in the Arabic text. The use of bagging further enhances robustness by reducing variance and mitigating overfitting, thus providing a solid benchmark against which transformer-based approaches can be evaluated. In addition to the gradient boosting baseline, we also evaluate a Multi-Layer Perceptron (MLP) with a single hidden layer.

4.4 Experimental Configuration

Our setup ensures rigorous evaluation through Stratified K-Fold cross-validation to maintain balanced class representation, the Optuna framework (Akiba et al., 2019) for Bayesian hyperparameter optimization, and macro-averaged Macro-F1 as the primary metric for balanced evaluation across classes. All transformer models are trained with GPU acceleration on an NVIDIA Tesla P100, ensuring efficient experimentation and reproducible results.

5 Results and Analysis

5.1 Baseline Performance

Baseline experiments without augmentation reveal important insights into Arabic hate and hope speech classification challenges. Ensemble bagging XGBoost achieved 0.626 Macro-F1, demonstrating traditional gradient boosting effectiveness with AraBERTv2 embeddings. MLP reached 0.616, showing marginal improvement despite neural architecture. Among individual Arabic BERT models, AraBERTv2 obtained 0.623, CAMeLBERT achieved 0.647, and MARBERT reached 0.639, representing comparable baseline performance with modest gains over traditional approaches.

These results reveal task characteristics: close performance between traditional ML and transformers suggests three-class classification challenges stem from inherent difficulties distinguishing between categories rather than sophisticated feature representations. Moderate Macro-F1s indicate significant challenges likely due to subtle distinctions and cultural context requirements.

Model	**Macro-F1**
Ensemble Bagging XGBoost	0.626
MLP	0.616
AraBERTv2	0.623
CAMeLBERT	0.647
MARBERT	0.639

Table 1: Baseline results without data augmentation (5-fold CV validation); Metric: Macro-F1

5.2 Impact of Data Augmentation and Ensemble Methods

Data augmentation yields substantial improvements across all models. AraBERTv2 improved to 0.665, MARBERT achieved 0.652, and CAMeLBERT reached 0.679. The hard voting ensemble

combining all three Arabic BERT variants achieved 0.689, representing the highest performance and demonstrating the effectiveness of leveraging complementary model strengths.

Consistent improvements validate our ensemble augmentation strategy combining back-translation and LLM-based few-shot generation. This approach addresses different data scarcity aspects while maintaining semantic properties essential for classification tasks.

Model	**Macro-F1**
Individual BERT Models	
AraBERTv2	0.665
MARBERT	0.652
CAMeLBERT	0.679
Ensemble Methods	
Hard Voting Ensemble	**0.689**
Ensemble Bagging XGBoost	0.656

Table 2: Results with data augmentation and ensemble methods (5-fold CV validation); Metric: Macro-F1

5.3 Model Analysis

The hard voting ensemble's superior performance (Macro-F1=0.689) demonstrates the value of combining diverse Arabic BERT variants, leveraging AraBERTv2's comprehensive coverage, CAMeLBERT's morphological optimization, and MARBERT's dialectal expertise. Among individual models, CAMeLBERT's strong performance (Macro-F1=0.679) stems from its optimized architecture for Arabic linguistic features.

Traditional ML competitive performance relative to individual transformers suggests primary challenges relate to dataset size and inherent task complexity rather than model capacity limitations, with practical implications for resource-limited scenarios.

6 Discussion

6.1 Task Complexity and Ensemble Benefits

The moderate Macro-F1 achieved by individual models (0.652–0.679) and the ensemble approach (0.687) underscore the inherent difficulty of three-class hate/hope speech classification, driven by subjective label boundaries, cultural context dependencies, and subtle linguistic cues. The superior results of the hard voting ensemble indicate that integrating diverse Arabic BERT variants can effectively leverage their complementary strengths to better handle these challenges.

6.2 Practical Implications

Ensemble data augmentation combining back-translation and LLM-based few-shot generation should be standard practice for Arabic datasets. Hard voting ensemble superiority reinforces the value of leveraging multiple Arabic-specific models over single architectures, with practical benefits for content moderation systems.

6.3 Future Directions

Promising directions include larger datasets with broader dialectal coverage, soft voting and weighted ensemble methods, multi-task learning approaches, and explainability research for ensemble decision-making processes.

7 Conclusion

We present an effective Arabic hate and hope speech classification approach using a hard voting ensemble of AraBERTv2, CAMeLBERT, and MARBERT with multi-hop back-translation and LLM-based few-shot augmentation, achieving an Macro-F1 of 0.689 in this competition. This work systematically evaluates leading Arabic models and validates that combining complementary architectures with advanced augmentation significantly boosts performance, providing a solid foundation for practical, culturally aware Arabic content moderation systems.

Acknowledgments

We thank anonymous reviewers for valuable feedback and the University of Information Technology for computational resources supporting this research.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Nguyen Minh Bao. 2025. baoflowin502 at MarsadLab: Baoflowin502 at mahed2025: Text-based hate and hope speech classification. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Ashraf Elnagar, Ridhwan Al-Debsi, and Omar Einea. 2020. Arabic text classification using deep learning models. *Information Processing & Management*, 57(1):102121.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. 2024. Datadream: Few-shot guided dataset generation. In *European Conference on Computer Vision*, pages 252–268. Springer.

Ahlam Wahdan, Mostafa Al-Emran, and Khaled Shaalan. 2024. A systematic review of arabic text classification: areas, applications, and future directions. *Soft Computing*, 28(2):1545–1566.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Georgios Mikros, Abul Hasnat, and Firoj Alam. 2025. MAHED shared task: Multimodal detection of hope and hate emotions in arabic content. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Wajdi Zaghouani, Hamdy Mubarak, and Md Rafiul Biswas. 2024. So hateful! building a multi-label hate speech annotated arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055.